

# Automatic Detection of Significant Areas for Functional Data with Directional Error Control

Xu, Peirong

Department of Mathematics, Southeast University, China and

Department of Statistics, Seoul National University, Korea

Lee, Youngjo

Department of Statistics, Seoul National University, Korea

Shi, Jian Qing \*

School of Mathematics & Statistics, Newcastle University, UK

**Abstract:** To detect differences between the mean curves of two samples in longitudinal study or functional data analysis, we usually need to partition the temporal or spatial domain into several pre-determined sub-areas. In this paper we apply the idea of large-scale multiple testing to find the significant sub-areas automatically in a general functional data analysis framework. A nonparametric Gaussian process regression model is introduced for two-sided multiple tests. We derive an optimal test which controls directional false discovery rates and propose a procedure by approximating it on a continuum. The proposed procedure controls directional false discovery rates at any specified level asymptotically. In addition, it is computationally inexpensive and able to accommodate different time points for observations across the samples. Simulation studies are presented to demonstrate its finite sample performance. We also apply it to an executive function research in children with Hemiplegic Cerebral Palsy and extend it to the equivalence tests.

*Key words:* False discovery rate; functional data; Gaussian process regression model; multiple testing; significant areas; Type III error.

---

\*Correspondence to: Dr J. Q. Shi, School of Mathematics & Statistics, Newcastle University, UK, j.q.shi@ncl.ac.uk.

# 1 Introduction

The testing problem in functional data analysis framework is motivated by an example on studying executive functions in children with Hemiplegic Cerebral Palsy. The Big/Little Circle (BLC) test is an attention measure that tests comprehension, learning and reversal of a rule (see e.g. Moore and Puri, 2012). In this study, the data on BLC mean correct latency was collected from 141 students, aging from 6 to 13, who completed the BLC test. Among them, 56% are action video game players (AVGPs) and 44% are non-action video game players (NAVGP) as shown in Figure 1. Let  $Y_1(t)$  and  $Y_2(t)$  be the BLC mean correct latency for NAVGP and AVGPs groups respectively, where  $t$  is the age of children. They are continuous functional variables although observations are collected at discrete points. We are interested in identifying ages that the means of  $Y_1(t)$  and  $Y_2(t)$  have significant difference. In particular, we wish to detect the specific areas of age where the significant differences occur. We will refer such areas as *significant areas*. Thus we wish to detect the significant areas automatically and at the same time minimize the false nondiscovery rate while controlling false discovery rates.

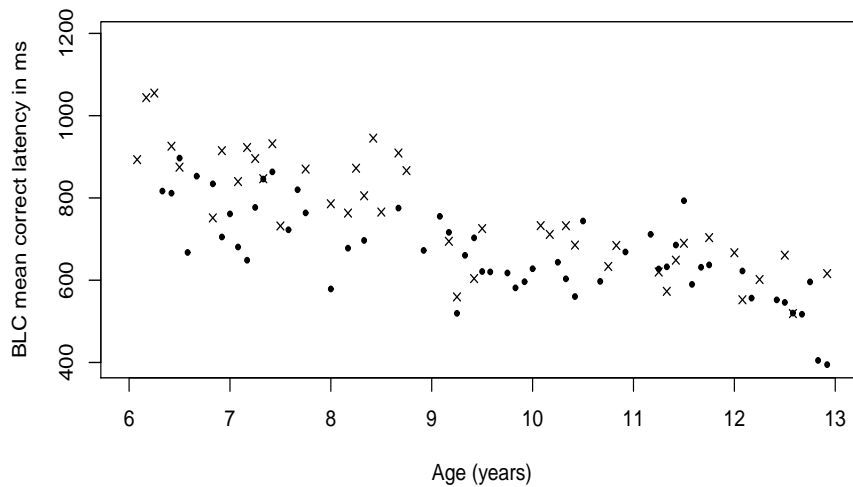


Figure 1: The scatterplot of BLC mean correct latency data in AVGPs group (.) and NAVGP group (x).

Functional data analysis (FDA) has emerged as a popular area of statistics over the last decade for the analysis of data with functional features, such as growth curves, motion and image data. Ramsay and Silverman (2005) and Ramsay et al. (2009) offered applied-oriented introductions to the ideas and tools of FDA. Ferraty and Romain (2011) reviewed some recent theoretical developments of FDA. Other important directions related to statistical inference in FDA includes Bosq (2000), Yao et al. (2005), Müller (2005), Ferraty and Vieu (2006), Di et al. (2009), Horváth and Kokoszka (2012), and Wang and Shi (2014) among many others. However,

hypothesis testing with directional error control on detecting areas in which differences of the mean curves of two samples are significant (i.e. detecting the *significant areas*) has received little attention. Inspired by the recent development of large-scale multiple testing for complex big data (see e.g. Zhang et al., 2011, Lee and Lee, 2014 and Sun et al. 2015), we propose an automatic detection procedure to find significant areas and allow control of the directional error at the same time.

Testing differences in the mean functions of two samples of curves has been approached in many literatures. For example, Zhang et al. (2010) introduced an  $L^2$ -norm based test, Horváth et al. (2013) developed a test based on the sample means of the curves, and Staicu et al. (2014) proposed a pseudo likelihood ratio test. Extension to multiple samples of curves was discussed in Shen and Faraway (2004). Cuevas et al. (2004), Estévez-Pérez and Vilar (2008) and Cuesta-Albertos and Febrero-Bande (2010) further extended it to the functional analysis of variance. Those works all focused on detecting the overall difference. However, we are often interested in determining the sub-areas of the functional domain (temporal or spatial) where the mean curves are significant different in many problems such as the motivating example we discussed earlier. To identify specific areas for a significant difference, Ramsay and Silverman (2005) proposed a pointwise t-test without multiplicity control, and Cox and Lee (2008) applied the Westfall-Young randomization method to control the family-wise error rate (FWER). However, when the number of null hypotheses is large, lack of multiplicity control is too permissive, while the full protection resulting from controlling the FWER is too stringent.

Compared with FWER in the context of multiple testing, the false discovery rate (FDR) introduced by Benjamini and Hochberg (1995) has received great attention during the past decade. Lots of procedures have been proposed in large-scale scientific studies with goals of controlling the FDR. For instance, Benjamini and Hochberg (1995) provided a sequential p-value method to control FDR; Sun and Cai (2009) introduced an asymptotical optimal procedure with test statistics under dependence; Liu et al. (2012) proposed a graphical-model based multiple testing procedure to genome-wide association studies; Lee and Bjørnstad (2013) expressed the problem of multiple testing as an inference problem with basic responses. Other relevant works are Storey (2002), Efron (2004, 2007), Genovese and Wasserman (2004), Zhang et al. (2011), French and Sain (2013) and some of the references therein. When the tests are two-sided as in our motivating example, it often becomes essential for researchers to further determine the direction of significance, rather than significance alone. Then, the decisions can potentially lead to three types of errors for each test: Type I error if the null hypothesis is true but rejected, Type II error if the null hypothesis is not true but failed to reject, and Type III error if the null hypothesis is not true but the direction of the alternative is falsely declared. To deal with Type I as well as Type III errors in the FDR framework, Benjamini and Yekutieli (2005) proposed a so-called directional Benjamini-Hochberg (BH) procedure for independent tests, Guo et al. (2010) extended the directional BH procedure based on the Bonferroni test to gene expression data with ordered categories, Clements et al. (2014) introduced a three-stage directional BH procedure to study vegetation fluctuations, and Lee and Lee (2014) developed an optimal extended likelihood

test with directional FDRs under hidden Markov random field models. However, the multiple testing problems mentioned above are all restricted to the assumption that each hypothesis has its own observed data, while in our motivating example, we only observed BLC mean current latency at finite time points in age range of [6, 13] but we need to make decisions at any age (time) between 6 and 13. Recently, Sun et al. (2015) developed an asymptotic optimal data-driven procedure that controls the FDR for multiple testing on a continuous domain, where the optimality is restricted in a set that test statistic satisfies monotone ratio condition. Their method is confined to change detection of one curve that may not be applicable to test differences in the two mean curves. And they derived the oracle procedure for two-sided tests by only controlling the FDR related to Type I error, which implies that their method may not be powerful in multiple tests with more than two actions.

To address the issue, we propose a new directional FDR procedure for detecting differences in the mean functions of two samples of functional data observed at discrete grid points. This would be the first attempt to handle two-sample multiple testing for detecting mean differences by controlling FDR in functional data analysis framework. In contrast to pointwise testing idea, we introduce a nonparametric Gaussian process regression model for directional two-sided multiple tests. It provides a natural framework on modeling mean structure and covariance structure of the difference between two curves simultaneously and the latter can be used to effectively extract information from nearby points for decision making. In the spirit of definitions in discrete cases, we define the directional FDRs for the continuous hypothesis testing process, and derive a test which optimally controls directional FDRs among all decision rules for multiple testing. Further, to make the continuous decision process applicable, a procedure is proposed by approximating the optimal test on a continuum. It is shown that it can control directional FDRs at any specified level asymptotically. Compared with conventional methods, our simulation studies manifest the drastically improved performance of the proposed procedure on directional error control and power.

The rest of the paper is organized as follows. In Section 2, we formulate the multiple testing problem and introduce directional FDRs for this continuous hypothesis testing process. Section 3 derives the optimal test with directional FDRs and presents a procedure for implementation. In Section 4, we investigate the finite sample performance of the proposed procedure by simulation studies and an application to the executive function study. The method is extended to equivalence tests in Section 5. The paper is concluded with a discussion in Section 6 and all the technical details are relegated to Appendix.

## 2 Problem formulation and directional FDRs

In this section, we formulate the multiple testing problem of detecting differences in the mean functions of two samples of curves and introduce directional FDRs on a continuum.

Let  $Y_g(t), g = 1, 2$  be two curves of functional data, which are functions of  $t$ . In functional

data analysis,  $t$  denotes a real-valued variable, which could be time or some other temporal or spatial variable. In this paper, without loss of generality, we assume  $t$  is time as in our motivating example and the corresponding time range is a closed interval  $T$ ; for simplicity take  $T = [0, 1]$ . We are interested in detecting differences between  $E(Y_1(t))$  and  $E(Y_2(t))$  over time on  $T$ . Specifically, consider the following functional regression model

$$\begin{aligned} Y_1(t) &= \mu(t) + \mu_d(t) + \epsilon_1(t), \\ Y_2(t) &= \mu(t) + \epsilon_2(t), \end{aligned} \tag{1}$$

where  $\mu(\cdot)$  and  $\mu_d(\cdot)$  are unknown functions and  $\epsilon_1(t)$  and  $\epsilon_2(t)$  are the independent random errors. Then, for each time  $t$ , we are interested in the directional two-sided test

$$\begin{aligned} &H_0(t) : |\mu_d(t)| \leq \Delta \\ \text{versus } &H_1(t) : \mu_d(t) < -\Delta \quad \text{or} \quad H_2(t) : \mu_d(t) > \Delta, \end{aligned} \tag{2}$$

where  $\Delta$  is a pre-specified constant, denoting the size of difference we are interested in. Assume that there is an underlying state  $z(t)$  associated with each time  $t$  taking one of three states. We set  $z(t) = 0$  if hypothesis at time  $t$  is the null and  $z(t) = 1$  or  $2$  if hypothesis at time  $t$  is the alternative 1 or 2, respectively. Let  $\delta(t) \in \{0, 1, 2\}$  be a decision rule for the hypothesis  $H_0(t)$ . If  $\delta(t) = z(t)$ , the hypothesis is correctly identified by the decision rule, otherwise there exist errors. Let  $R_k = \{t \in T : \delta(t) = k\}$  and  $V_{jk} = \{t \in T : z(t) = j, \delta(t) = k\}$  for  $j, k = 0, 1, 2$ . Table 1 summarizes the possible outcomes of multiple testing with two alternatives, which shows that there exist three types of errors in the directional two-sided multiple testing (2).

Table 1: Outcomes of multiple testing with two alternatives

	Declared as null $\delta(t) = 0$	Declared as alternative 1 $\delta(t) = 1$	Declared as alternative 2 $\delta(t) = 2$	Total
Null ( $z(t) = 0$ )	$V_{00}$	$V_{01}$ ( <i>Type I error</i> )	$V_{02}$ ( <i>Type I error</i> )	$T_0$
Alternative 1 ( $z(t) = 1$ )	$V_{10}$ ( <i>Type II error</i> )	$V_{11}$	$V_{12}$ ( <i>Type III error</i> )	$T_1$
Alternative 2 ( $z(t) = 2$ )	$V_{20}$ ( <i>Type II error</i> )	$V_{21}$ ( <i>Type III error</i> )	$V_{22}$	$T_2$
Total	$R_0$	$R_1$	$R_2$	$T$

Let  $\mathcal{L}(\cdot)$  be the Lebesgue measure on time range  $T$ . Then,  $\mathcal{L}(N_1) = \mathcal{L}(V_{01}) + \mathcal{L}(V_{02})$  and  $\mathcal{L}(N_2) = \mathcal{L}(V_{10}) + \mathcal{L}(V_{20})$  are the sizes of areas corresponding to Type I and Type II errors, respectively, and  $\mathcal{L}(N_3) = \mathcal{L}(V_{12}) + \mathcal{L}(V_{21})$  is the size of area corresponding to Type III error, a directional error. When the interest is to test hypotheses at individual time points, a natural and practical way is to control an error rate in the FDR framework by considering all of these three types of errors. Thus, in this paper, we propose to control either the sum of Type I and

Type III errors while minimizing the Type II error or control the Type I error while minimizing the sum of Type II and Type III errors. Let  $a \vee b = \max\{a, b\}$ . Define FDR and the marginal FDR (mFDR) for Type I error rate as

$$\text{FDR}_I = E \left\{ \frac{\mathcal{L}(N_1)}{\mathcal{L}(R_1 \cup R_2) \vee 1} \right\} \text{ and } \text{mFDR}_I = \frac{E\{\mathcal{L}(N_1)\}}{E\{\mathcal{L}(R_1 \cup R_2)\}},$$

those for Type III error rate as

$$\text{FDR}_{III} = E \left\{ \frac{\mathcal{L}(N_3)}{\mathcal{L}(R_1 \cup R_2) \vee 1} \right\} \text{ and } \text{mFDR}_{III} = \frac{E\{\mathcal{L}(N_3)\}}{E\{\mathcal{L}(R_1 \cup R_2)\}},$$

and those for the sum of the Type I and Type III error rates as

$$\text{FDR}_{I+III} = E \left\{ \frac{\mathcal{L}(N_1 \cup N_3)}{\mathcal{L}(R_1 \cup R_2) \vee 1} \right\} \text{ and } \text{mFDR}_{I+III} = \frac{E\{\mathcal{L}(N_1 \cup N_3)\}}{E\{\mathcal{L}(R_1 \cup R_2)\}}.$$

Besides the error rate for discoveries, we can define similar error rate for false nondiscoveries, the false nondiscovery rate and the marginal false nondiscovery rate

$$\text{FNDR} = E \left\{ \frac{\mathcal{L}(N_2)}{\mathcal{L}(R_0) \vee 1} \right\} \text{ and } \text{mFNDR} = \frac{E\{\mathcal{L}(N_2)\}}{E\{\mathcal{L}(R_0)\}},$$

which is related to Type II error. Further, to compute the power of a single directional two-sided testing procedure, Leventhal and Huynh (1996) recommended excluding Type III error from the conventional power. Therefore, in this paper, we define a modified power (MP) of a directional two-sided multiple testing procedure by considering both Type II and Type III errors

$$\text{MP} = 1 - \frac{E\{\mathcal{L}(N_2 \cup N_3)\}}{E\{\mathcal{L}(T_1 \cup T_2)\}}.$$

REMARK 1. *Lee and Lee (2014) created a similar table to summarize the outcomes of multiple testing with two alternatives and defined the corresponding directional FDRs. The key difference here is that for continuous testing process (2), the false discovery measures are related to the sizes of areas corresponding to three types of errors, which couldn't be calculated directly by counting the number of cases as in discrete case where each hypothesis has its own observed data. Therefore, a new strategy is needed to develop for inference based on the continuous functional data analysis framework but using the data observed at discrete points.*

### 3 Optimal tests for automatic detection of significant areas

#### 3.1 Optimal procedures for controlling directional FDRs

Suppose the observed data  $\{(Y_{1i}, t_{1i}) : i = 1, \dots, n_1\}$  and  $\{(Y_{2i}, t_{2i}) : i = 1, \dots, n_2\}$  are realizations of two underlying stochastic processes from model (1). The notation of the time points,

$t_{1i}$  and  $t_{2i}$ , allows for different observation points in the two groups, and  $t_{1i}$ 's and  $t_{2i}$ 's consist of subsets of  $T$ . Our objective is to predict the states of hypothesis  $z(t) \in \{0, 1, 2\}$  at any time point  $t \in T$  in an optimal way. Therefore, it is necessary to exploit the temporal correlations and extract information from nearby points for prediction. Consider a loss function

$$L(\delta, z; \lambda) = \lambda_1 \mathcal{L}(N_1) + \lambda_2 \mathcal{L}(N_2) + \lambda_3 \mathcal{L}(N_3), \quad (3)$$

where  $\lambda_1, \lambda_2$  and  $\lambda_3$  are relative costs. The following theorem derives the optimal rule for the weighted classification problem (3).

**THEOREM 1.** *Let  $\mathcal{D}$  be the whole data set consisting of  $\{(Y_{1i}, t_{1i}) : i = 1, \dots, n_1\}$  and  $\{(Y_{2i}, t_{2i}) : i = 1, \dots, n_2\}$ . Assume all parameters in model (1) are known. Then,*

(1) *if  $\lambda_2 = 1$  and  $\lambda_1 = \lambda_3 = \lambda$  in (3), the optimal decision rule  $\delta^{(I+III)} = \{\delta^{(I+III)}(t) : t \in T\} = \operatorname{argmin}_{\delta} E\{L(\delta, z; \lambda) | \mathcal{D}\}$  becomes*

$$\begin{aligned} \delta^{(I+III)}(t) &= 2 \text{ if } \frac{1 - P(z(t) = 0 | \mathcal{D})}{1 - P(z(t) = 2 | \mathcal{D})} > \lambda \text{ and } P(z(t) = 2 | \mathcal{D}) > P(z(t) = 1 | \mathcal{D}), \\ &= 1 \text{ if } \frac{1 - P(z(t) = 0 | \mathcal{D})}{1 - P(z(t) = 1 | \mathcal{D})} > \lambda \text{ and } P(z(t) = 2 | \mathcal{D}) \leq P(z(t) = 1 | \mathcal{D}), \\ &= 0 \text{ otherwise;} \end{aligned}$$

(2) *if  $\lambda_1 = \lambda$  and  $\lambda_2 = \lambda_3 = 1$  in (3), the optimal decision rule  $\delta^{(I)} = \{\delta^{(I)}(t) : t \in T\} = \operatorname{argmin}_{\delta} E\{L(\delta, z; \lambda) | \mathcal{D}\}$  becomes*

$$\begin{aligned} \delta^{(I)}(t) &= 2 \text{ if } \frac{P(z(t) = 2 | \mathcal{D})}{P(z(t) = 0 | \mathcal{D})} > \lambda \text{ and } P(z(t) = 2 | \mathcal{D}) > P(z(t) = 1 | \mathcal{D}), \\ &= 1 \text{ if } \frac{P(z(t) = 1 | \mathcal{D})}{P(z(t) = 0 | \mathcal{D})} > \lambda \text{ and } P(z(t) = 2 | \mathcal{D}) \leq P(z(t) = 1 | \mathcal{D}), \\ &= 0 \text{ otherwise.} \end{aligned}$$

Theorem 1 gives the optimal rules for various weighted classification problems. We next show that the optimality property can be extended to the multiple testing problems with respect to various directional FDRs defined in Section 2.

**THEOREM 2.** *Let  $\mathcal{A} = \{\delta^{(I+III)} : \lambda > 0\}$  be the collection of decision rules in form of  $\delta^{(I+III)}$  derived in Theorem 1. Given an  $m\text{FDR}_{I+III}$  level  $\alpha$ , let  $\delta = \{\delta(t) : t \in T\}$  be any decision rule satisfying  $m\text{FDR}_{I+III}\{\delta\} \leq \alpha$ . Then, there exists a  $\lambda$  determined by  $\delta$  such that  $\delta^{(I+III)} \in \mathcal{A}$  performs better than  $\delta$  in the sense that*

$$m\text{FDR}_{I+III}\{\delta^{(I+III)}\} \leq m\text{FDR}_{I+III}\{\delta\} \leq \alpha,$$

and

$$m\text{FNDR}\{\delta^{(I+III)}\} \leq m\text{FNDR}\{\delta\}.$$

Theorem 2 demonstrates that the optimal decision rule for controlling the sum of Type I and Type III errors with the smallest Type II error belongs to the set  $\mathcal{A}$ . In other words, one only needs to search in  $\mathcal{A}$  for the optimal rule, instead of searching for all decision rules. Similarly, it can be shown that the optimal decision rule for controlling Type I error with the smallest sum of Type II and Type III errors is in the form of  $\delta^{(I)}$  derived in Theorem 1.

REMARK 2. *In the case of  $\lambda_3 = 0$ , Sun et al. (2015) showed the optimal solution to the weighted classification problem is optimal in  $\{\delta : \delta(t) = I\{T(t) < c\}, T \text{ satisfies monotone ratio condition}\}$  for the multiple testing problem, but Theorem 2 extends the result to a more general case, revealing that this solution is even optimal among all decision rules for the multiple testing.*

### 3.2 Extension to practical situations

It is not straightforward to use the optimal procedures described in Section 3.1 because (a) it is impossible to make an uncountable number of decisions on  $T$ , and (b) the true smooth trajectories  $\mu(\cdot)$  and  $\mu_d(\cdot)$  are not directly observable and thus the test statistics should be evaluated at unobserved time points. In this section, we develop procedures for directional FDRs control to overcome these difficulties.

To address (a), we first divide the interval  $T = [0, 1]$  into  $N$  equal-length subintervals  $[s_{i-1}, s_i]$  with  $s_0 = 0$  and  $s_i = s_{i-1} + 1/N$ ,  $i = 1, \dots, N-1$ , and pick the center point  $t_i^*$  in  $[s_{i-1}, s_i]$ ,  $i = 1, \dots, N-1$ . Then, for a decision rule  $\delta$ , we have

$$\begin{aligned} E\{\mathcal{L}(N_1)\} &= \int_0^1 E\{I(\delta(t) \neq 0)P(z(t) = 0|\mathcal{D})\}d\mathcal{L}(t) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N E\{S_0(t_i^*)I(\delta(t_i^*) \neq 0)\}, \\ E\{\mathcal{L}(N_3)\} &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N E\{S_2(t_i^*)I(\delta(t_i^*) = 1) + S_1(t_i^*)I(\delta(t_i^*) = 2)\}, \text{ and} \\ E\{\mathcal{L}(N_1 \cup N_3)\} &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^2 \sum_{j \neq k} E\{S_j(t_i^*)I(\delta(t_i^*) = k)\} \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^2 E\{I(\delta(t_i^*) = k)(1 - S_k(t_i^*))\}, \end{aligned}$$

where function  $S_k(t) = P(z(t) = k|\mathcal{D})$ ,  $k = 0, 1, 2$ . Therefore, motivated by the limit definition of a definite integral, mFDR<sub>I</sub> can be estimated by

$$\widehat{\text{mFDR}}_I(\lambda) = \frac{1}{r} \sum_{i=1}^N S_0(t_i^*)I(\delta(t_i^*) \neq 0) \quad (4)$$

for any given  $\lambda$  and all parameters in model (1), where  $r = \sum_{i=1}^N I(\delta(t_i^*) \neq 0)$ . According to Theorem 1, it is easy to see that  $\delta^{(I)}(t) = 0$  if  $S_0(t) \leq (1 + \lambda)^{-1}$ . Suppose that  $\lambda_1$  and  $\lambda_2$  are chosen so that  $h_{(r)} < (1 + \lambda_2)^{-1} < h_{(r+1)} < (1 + \lambda_1)^{-1} < h_{(r+2)}$ , where  $h_{(r)}$  is the  $r$ th smallest



value of  $S_0(t_i^*)$ . Then,

$$\begin{aligned}\widehat{\text{mFDR}}_{\text{I}}(\lambda_2) - \widehat{\text{mFDR}}_{\text{I}}(\lambda_1) &= r^{-1} \sum_{i=1}^r h_{(i)} - (r+1)^{-1} \sum_{i=1}^{r+1} h_{(i)} \\ &= \{r(r+1)\}^{-1} \left\{ \sum_{i=1}^r h_{(i)} - r h_{(r+1)} \right\} < 0.\end{aligned}$$

Thus,  $\widehat{\text{mFDR}}_{\text{I}}$  monotonically decreases with  $\lambda$ , and we propose the following step-down test procedure for  $\text{FDR}_{\text{I}}$  control:

$$\begin{aligned}\text{let } \lambda^* &= \inf\{\lambda : \widehat{\text{mFDR}}_{\text{I}}(\lambda) \leq \alpha\}; \text{ then} \\ \delta^{(I)}(t) &= \sum_{i=1}^N I(s_{i-1} \leq t < s_i) \delta^{(I)}(t_i^*) \\ \text{with } \delta^{(I)}(t_i^*) &= 2 \text{ if } \frac{S_2(t_i^*)}{S_0(t_i^*)} > \lambda^* \text{ and } S_2(t_i^*) > S_1(t_i^*), \\ &= 1 \text{ if } \frac{S_1(t_i^*)}{S_0(t_i^*)} > \lambda^* \text{ and } S_2(t_i^*) \leq S_1(t_i^*), \\ &= 0 \text{ otherwise.}\end{aligned} \tag{5}$$

The following theorem shows that this test controls  $\text{FDR}_{\text{I}}$  at level  $\alpha$  asymptotically, which implies that the proposed procedure (5) approximates a multiple comparison correction for a continuous comparison process (2) as the grid for pointwise comparisons becomes finer.

**THEOREM 3.** *Let  $\{\cup_{i=1}^N [s_{i-1}, s_i) : N = 1, 2, \dots\}$  be a sequence of partitions of  $T$  satisfying Conditions C1 and C2 in the Appendix. Then, the  $\text{FDR}_{\text{I}}$  level of procedure (5) satisfies  $\text{FDR}_{\text{I}} \leq \alpha + o(1)$  when  $N \rightarrow \infty$ .*

**REMARK 3.** *For simplicity, we choose the center point  $t_i^*$  in each subinterval  $[s_{i-1}, s_i)$  as a representative point. But from the proof of Theorem 3, we can see that, no matter which point is chosen as a representative point in  $[s_{i-1}, s_i)$ , the proposed procedure (5) controls  $\text{FDR}_{\text{I}}$  at the nominal level asymptotically as long as Conditions C1 and C2 are fulfilled.*

Similarly, by using  $\widehat{\text{mFDR}}_{\text{I+III}}(\lambda) = \frac{1}{r} \sum_{i=1}^N \sum_{k=1}^2 I(\delta(t_i^*) = k)(1 - S_k(t_i^*))$ , we control  $\text{FDR}_{\text{I+III}}$  at the nominal level. However, they are still difficult to implement because of (b).

Further to address (b), we propose a Gaussian process regression (GPR) model for (1) to estimate unknown quantities  $S_k(t) = P(z(t) = k | \mathcal{D})$ ,  $k = 0, 1, 2$ . GPR model is a good choice as a globally approximated nonlinear functional regression model in (1) (in contrast with locally approximated model for most of conventional nonparametric model); see the details in Shi and Choi (2011) and Wang and Shi (2014). Specifically, consider  $\{\mu(t) : t \in T\}$  and  $\{\mu_d(t) : t \in T\}$  as independent random processes and suppose they have Gaussian process priors with zero means and kernel functions  $\kappa(\cdot, \cdot; \boldsymbol{\eta})$  and  $\gamma(\cdot, \cdot; \boldsymbol{\theta})$ , respectively, where  $\text{Cov}(\mu(t), \mu(t')) = \kappa(t, t'; \boldsymbol{\eta})$  and  $\text{Cov}(\mu_d(t), \mu_d(t')) = \gamma(t, t'; \boldsymbol{\theta})$ . Assume that  $\{\epsilon_1(t) : t \in T\}$  and  $\{\epsilon_2(t) : t \in T\}$  are Gaussian white noise processes with zero mean and variance  $\sigma^2$ , which are independent from each other

and to both  $\{\mu(t) : t \in T\}$  and  $\{\mu_d(t) : t \in T\}$ . One example of the kernel function  $\gamma(\cdot, \cdot; \boldsymbol{\theta})$  is the following squared exponential covariance function with a nonstationary linear term:

$$\gamma(t_i, t_j; \boldsymbol{\theta}) = \xi \exp \left\{ -\omega(t_i - t_j)^2 / 2 \right\} + \zeta t_i t_j, \quad (6)$$

where  $\boldsymbol{\theta} = (\xi, \omega, \zeta)$  is a set of hyper-parameters. When  $\zeta = 0$ , the kernel function  $\gamma(\cdot, \cdot; \boldsymbol{\theta})$  reduces to so-called squared exponential covariance function, which is stationary and nondegenerate (Rasmussen and Williams, 2006). The parameter  $\omega$  corresponds to the smoothing parameters in spline. So, we call  $\omega^{-1}$  the length-scale. A large length-scale implies the underlying curve is expected to be essentially flat and the decrease in length-scale results in more rapidly fluctuating functions.

Let  $n = n_1 + n_2$ ,  $\boldsymbol{\Theta} = (\boldsymbol{\eta}^T, \boldsymbol{\theta}^T, \sigma^2)^T$ , and  $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T)^T$  with  $\mathbf{Y}_1 = (Y_{11}, \dots, Y_{1n_1})^T$  and  $\mathbf{Y}_2 = (Y_{21}, \dots, Y_{2n_2})^T$ . Consider the joint density function of  $\mathbf{Y}, \tilde{\boldsymbol{\mu}}$  and  $\tilde{\boldsymbol{\mu}}_d$

$$f_{\boldsymbol{\Theta}}(\mathbf{Y}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}_d) = \phi(\tilde{\boldsymbol{\mu}} \mid \mathbf{0}, \mathbf{K}_n) \phi(\tilde{\boldsymbol{\mu}}_d \mid \mathbf{0}, \boldsymbol{\Gamma}_{n_1}) \prod_{i=1}^{n_1} \phi(Y_{1i} \mid \mu(t_{1i}) + \mu_d(t_{1i}), \sigma^2) \prod_{i=1}^{n_2} \phi(Y_{2i} \mid \mu(t_{2i}), \sigma^2), \quad (7)$$

where  $\tilde{\boldsymbol{\mu}} = (\mu(t_{11}), \dots, \mu(t_{1n_1}), \mu(t_{21}), \dots, \mu(t_{2n_2}))^T$ ,  $\tilde{\boldsymbol{\mu}}_d = (\mu_d(t_{21}), \dots, \mu_d(t_{2n_2}))^T$ ,  $\phi(\cdot)$  is the density of (multivariate) normal distribution,  $\mathbf{K}_n$  and  $\boldsymbol{\Gamma}_{n_1}$  are covariance matrices of  $\tilde{\boldsymbol{\mu}}$  and  $\tilde{\boldsymbol{\mu}}_d$ , respectively, with  $(i, j)$ th element  $\kappa(t_i, t_j; \boldsymbol{\eta})$  and  $\gamma(t_i, t_j; \boldsymbol{\theta})$ . Then, the parameters  $\boldsymbol{\Theta}$  can be consistently estimated by maximizing the likelihood (see Shi and Choi, 2011)

$$l(\boldsymbol{\Theta}; \mathbf{Y}) = f_{\boldsymbol{\Theta}}(\mathbf{Y}) = \int \int f_{\boldsymbol{\Theta}}(\mathbf{Y}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}_d) d\tilde{\boldsymbol{\mu}} d\tilde{\boldsymbol{\mu}}_d.$$

Let  $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\eta}}^T, \hat{\boldsymbol{\theta}}^T, \hat{\sigma}^2)^T$  be the estimates of  $\boldsymbol{\Theta}$ . Then, we can make inference about  $\mu_d(t)$  by using  $f_{\hat{\boldsymbol{\Theta}}}(\mu_d(t) \mid \mathcal{D})$ . As the sample size  $n$  goes to infinity, we have  $f_{\hat{\boldsymbol{\Theta}}}(\mu_d(t) \mid \mathcal{D}) \rightarrow f_{\boldsymbol{\Theta}}(\mu_d(t) \mid \mathcal{D})$ .

Now, we consider how to make inference about  $\boldsymbol{\mu}_d^N = (\mu_d(t_1^*), \dots, \mu_d(t_N^*))^T$ , where  $\mathbf{T}^* = (t_1^*, \dots, t_N^*)$  is a collection of the center points based on partition  $T = \cup_{i=1}^N [s_{i-1}, s_i]$ . It is not difficult to prove (see the details in Appendix B) that the conditional distribution of  $\boldsymbol{\mu}_d^N$  given the data set  $\mathcal{D}$  is a multivariate normal distribution with mean and covariance given by

$$\begin{aligned} \bar{\boldsymbol{\mu}} &\equiv \mathbb{E}(\boldsymbol{\mu}_d^N \mid \mathcal{D}) = \boldsymbol{\Psi}(\mathbf{T}^*) \{ \sigma^2 \mathbf{I}_{n_1} + (\mathbf{I}_{n_1} - \boldsymbol{\Sigma}_{11}) \boldsymbol{\Gamma}_{n_1} \}^{-1} \{ (\mathbf{I}_{n_1} - \boldsymbol{\Sigma}_{11}) \mathbf{Y}_1 - \boldsymbol{\Sigma}_{12} \mathbf{Y}_2 \}, \\ \boldsymbol{\Lambda} &\equiv \text{Cov}(\boldsymbol{\mu}_d^N \mid \mathcal{D}) = \boldsymbol{\Gamma}_N - \boldsymbol{\Psi}(\mathbf{T}^*) \boldsymbol{\Gamma}_{n_1}^{-1} \boldsymbol{\Psi}^T(\mathbf{T}^*) + \sigma^2 \boldsymbol{\Psi}(\mathbf{T}^*) \boldsymbol{\Omega}_{n_1}^{-1} \boldsymbol{\Psi}^T(\mathbf{T}^*), \end{aligned} \quad (8)$$

where  $\boldsymbol{\Psi}(\mathbf{T}^*)$  is the  $N \times n_1$  covariance matrix between  $\boldsymbol{\mu}_d^N$  and  $\tilde{\boldsymbol{\mu}}_d$  with  $(i, j)$ th element  $\gamma(t_i^*, t_j; \boldsymbol{\theta})$ ,  $\boldsymbol{\Gamma}_N$  is the covariance matrix of  $\boldsymbol{\mu}_d^N$  with  $(i, j)$ th element  $\gamma(t_i^*, t_j^*; \boldsymbol{\theta})$ ,  $\boldsymbol{\Sigma}$  is a  $n \times n$  block matrix given by

$$\boldsymbol{\Sigma} = \mathbf{K}_n (\mathbf{K}_n + \sigma^2 \mathbf{I}_n)^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

with  $\mathbf{I}_n$  being a  $n \times n$  identity matrix, and  $\boldsymbol{\Omega}_{n_1} = \sigma^2 \boldsymbol{\Gamma}_{n_1} + \boldsymbol{\Gamma}_{n_1} (\mathbf{I}_{n_1} - \boldsymbol{\Sigma}_{11}) \boldsymbol{\Gamma}_{n_1}$ . Therefore, to calculate  $\widehat{\text{mFDR}}_1$  define in (4), we draw  $M$  samples  $\{\hat{\boldsymbol{\mu}}_m^N : m = 1, \dots, M\}$  from the conditional distribution of  $\boldsymbol{\mu}_d^N = (\mu_d(t_1^*), \dots, \mu_d(t_N^*))^T$ , where  $\hat{\boldsymbol{\mu}}_m^N = (\hat{\mu}_{m1}^N, \dots, \hat{\mu}_{mN}^N)^T$  is the  $m$ th

$N$ -dimensional sample predicting the values at time points  $t_1^*, \dots, t_N^*$ . Then, we can approximate  $\widehat{\text{mFDR}}_1$  by replacing  $S_0(t_i^*)$  by its estimate  $\widehat{S}_0(t_i^*)$ . More specifically, note that

$$\begin{aligned} S_0(t_i^*) &= P(z(t_i^*) = 0 \mid \mathcal{D}) = E[I\{|\mu_d(t_i^*)| \leq \Delta\} \mid \mathcal{D}] \\ &= \int I\{|\mu_d(t_i^*)| \leq \Delta\} \phi(\boldsymbol{\mu}_d^N \mid \bar{\boldsymbol{\mu}}, \boldsymbol{\Lambda}) d\boldsymbol{\mu}_d^N. \end{aligned}$$

Thus,  $S_0(t_i^*)$  can be estimated by

$$\widehat{S}_0(t_i^*) = \frac{1}{M} \sum_{m=1}^M I\{|\widehat{\mu}_{mi}^N| \leq \Delta\}.$$

Similarly, to implement procedure (5), we compute  $S_1(t_i^*)$  and  $S_2(t_i^*)$  by

$$\begin{aligned} \widehat{S}_1(t_i^*) &= \frac{1}{M} \sum_{m=1}^M I\{\widehat{\mu}_{mi}^N < -\Delta\} \\ \text{and } \widehat{S}_2(t_i^*) &= \frac{1}{M} \sum_{m=1}^M I\{\widehat{\mu}_{mi}^N > \Delta\}, \end{aligned}$$

respectively.

REMARK 4. *The joint density function defined in (7) is the  $h$ -likelihood (Lee and Nelder, 1996) when we treat  $\bar{\boldsymbol{\mu}}$  and  $\bar{\boldsymbol{\mu}}_d$  as unobservable random variables. It contains all the information in the data for parameters  $\boldsymbol{\Theta}$  and unobservable random variables  $\bar{\boldsymbol{\mu}}$  and  $\bar{\boldsymbol{\mu}}_d$  (Bjørnstad, 1996). The method discussed above can also be extended to a fully Bayesian way by assuming a hyper-prior distribution for  $\boldsymbol{\Theta}$ ; see Shi and Choi (2011).*

REMARK 5. *Sun et al. (2015) used the similar approximation strategy to mimic the optimal procedure as in (5). But for implementation, they applied a Bayesian computational algorithm and drew MCMC samples during the iterations to estimate  $S_0(t_i^*)$ , which can be rather computationally intensive when the number of representative points  $N$  is large. While for the proposed procedure, we can get the estimates of unknown parameters efficiently by using the nice proprieties of GPR models and estimate  $S_k(t_i^*), k = 0, 1, 2$  directly by generating the samples from multivariate normal distribution with mean and covariance given in (8). GPR models can cope with multiple covariates and thus the proposed method can be easily extend to problems in multivariate functional domain for example in 3-dimensional spatial domain or 4-dimensional temporal/spatial domain dynamical fMRI images.*

## 4 Numerical study

### 4.1 Simulation studies

In this subsection, we conduct a set of simulation studies to assess the finite sample performance of the proposed method. The purpose is twofold. First, we compare our method with directional

Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) and directional Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001, 2005). Since both of them only work for discrete case where each hypothesis has its own observed data, in Example 1, we assume two curves observed at the same set of time points and restrict the analysis for testing hypotheses at this set to permit comparisons, which means we have  $n_1 = n_2 = N$ . Second, we evaluate the performance of our method in Example 2 to test hypotheses on a continuum  $T$  with two curves observed at different discrete grid points.

**EXAMPLE 1** We generate 200 datasets from model (1), where both  $\mu(\cdot)$  and  $\mu_d(\cdot)$  are Gaussian processes with zero means and covariance functions  $\kappa(t_i, t_j; \boldsymbol{\eta}) = 3 \exp\{-(t_i - t_j)^2\}$  and  $\gamma(t_i, t_j; \boldsymbol{\theta}) = 10 \exp\{-\omega(t_i - t_j)^2/2\}$ , respectively, implying two stationery processes, and the error processes  $\epsilon_1(\cdot)$  and  $\epsilon_2(\cdot)$  are white noise processes with zero mean and finite variance  $\sigma^2 = 1$ . For each simulated dataset, data are generated at  $N = 500$  time points  $t_i \sim \text{Uniform}([6, 13])$ . For all simulations, we choose  $\Delta = 0.80$  so that the expected proportion of time points with  $|\mu_d(t)| \leq \Delta$  is 20% and set the nominal level as  $\alpha = 0.10$ . To study the effects of correlation, we vary  $\omega$  resulting in the curve  $\mu_d(\cdot)$  from smooth to fluctuating.

Figure 2 plots  $\text{FDR}_{\text{I+III}}$  and  $\text{FDR}_{\text{I}}$  as functions of  $\omega$  at the nominal level 0.10 and Figure 3 shows the averages of FNDR and MP over the 200 datasets. We can see that the proposed method control  $\text{FDR}_{\text{I}}$  and  $\text{FDR}_{\text{I+III}}$  reasonably well. When  $\omega$  becomes larger, there is a increasing chance to detect  $\mu_d(t)$  to be non-null and declare it to be less than  $-\Delta$  or larger than  $\Delta$ . That is, as the correlation of the signals decaying, it is more possible to make directional errors. As expected, Figure 3(a) shows that the proposed method may have relatively large FNDR when  $\omega$  is quite large. Correspondingly, Figure 3(b) implies that it may encounter loss of power. Though the directional Benjamini-Yekutieli procedure accounts for dependence, it is the most conservative and therefore the least powerful. The directional Benjamini-Hochberg procedure, derived under the independence assumption, controls the  $\text{FDR}_{\text{I}}$  conservatively as the original Benjamini and Hochberg's procedure (Benjamini and Hochberg, 1995), which controls the  $\text{FDR}_{\text{I}}$  at a level smaller than the desired  $\alpha$ .

**EXAMPLE 2** In this example, the true model is the same as in Example 1, except that the process  $\mu(\cdot)$  is a Gaussian process with zero mean but a nonstationary covariance function  $\kappa(t_i, t_j; \boldsymbol{\eta}) = 3 \exp\{-(t_i - t_j)^2\} + 3t_i t_j$ . The sampling design for two curves is balanced ( $n_1 = n_2 = 200$ ), but irregular, and furthermore different across the two samples. Specifically, we assume that  $\{t_{1i} : i = 1, \dots, n_1\}$  and  $\{t_{2i} : i = 1, \dots, n_2\}$  are iid realizations from  $\text{Uniform}([6, 13])$  with 20% overlapping. Predictions are made and tests of (2) are conducted at center points of  $N = 500$  equal-length subintervals covering the time range  $[6, 13]$ . For all simulations, we set  $\Delta = 0.80$ ,  $\alpha = 0.10$  and vary the value of  $\omega$  as in Example 1 and repeat our procedure 200 times for each configuration.

Figure 4 depicts the distribution of  $\text{FDR}_{\text{I+III}}$  and  $\text{FDR}_{\text{I}}$  and Figure 5 presents the distribution of FNDR and MP over 200 replications. We can see that the proposed method maintains  $\text{FDR}_{\text{I+III}}$  and  $\text{FDR}_{\text{I}}$  properly no matter the curve  $\mu_d(\cdot)$  is smooth or wiggly. It is in accordance with Theorem 3 in Section 3.2. And it implies that the proposed procedure is robust under

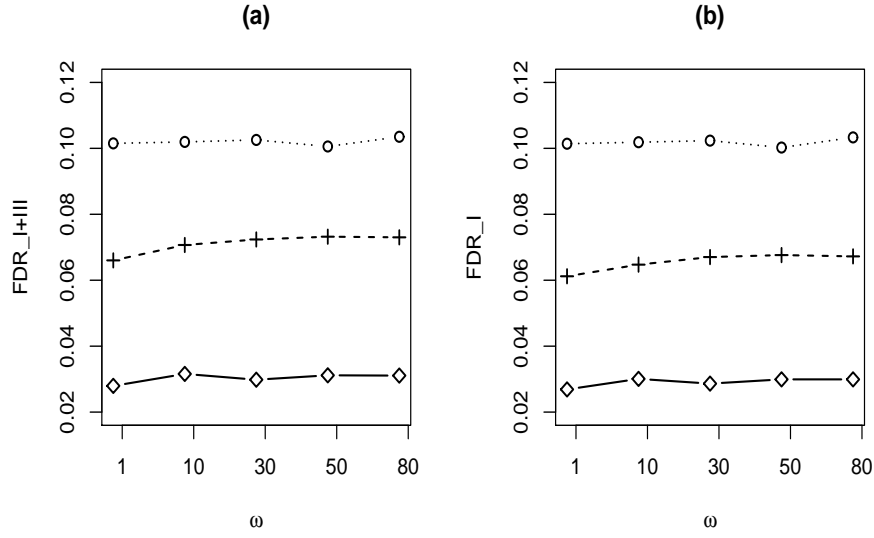


Figure 2: Comparison of directional Benjamini-Hochberg procedure (+), directional Benjamini-Yekutieli procedure (◇) and the proposed method (o): (a)  $FDR_{I+III}$  versus  $\omega$ ; (b)  $FDR_I$  versus  $\omega$ .

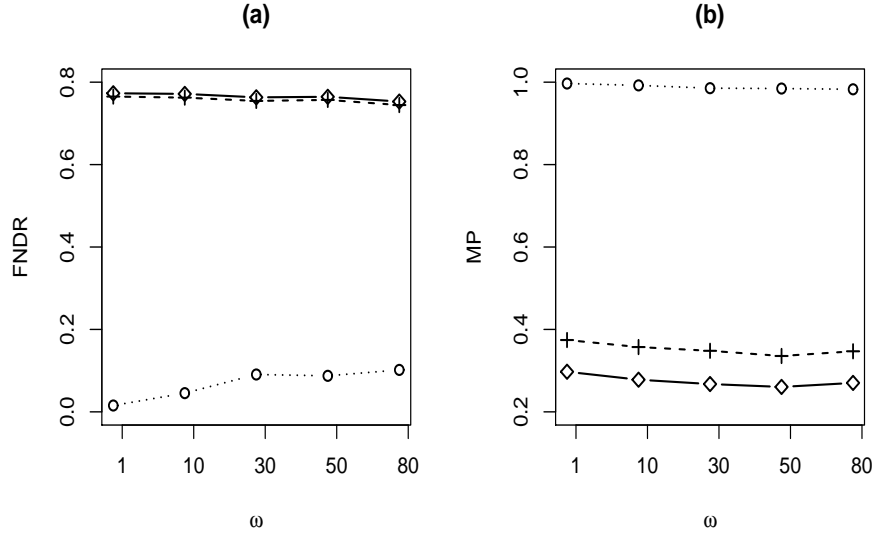


Figure 3: Comparison of directional Benjamini-Hochberg procedure (+), directional Benjamini-Yekutieli procedure (◇) and the proposed method (o): (a) FNDR versus  $\omega$  under  $FDR_{I+III}$  at 0.10; (b) MP versus  $\omega$  under  $FDR_I$  at 0.10.

different magnitudes of dependence across the values of  $\mu_d(\cdot)$ . Moreover, the boxplots of FNDR and MP show that the proposed procedure is powerful, where the MP is 0.95 even when  $\omega$  is very large.

To investigate the consistency of the estimated directional errors, the values of the estimated  $E\{\mathcal{L}(N_1)\}$ ,  $E\{\mathcal{L}(N_2)\}$ ,  $E\{\mathcal{L}(N_3)\}$ ,  $E\{\mathcal{L}(R_0)\}$ ,  $E\{\mathcal{L}(R_1)\}$  and  $E\{\mathcal{L}(R_2)\}$  are averaged so that the directional errors are calculated and regarded as the true values. Table 2 compares them with the estimated directional errors when  $\omega = 80$ . We observe that the proposed procedure gives consistent estimators. Slight underestimation of mFDR explains slightly liberal control of directional FDRs when  $\omega$  is large.

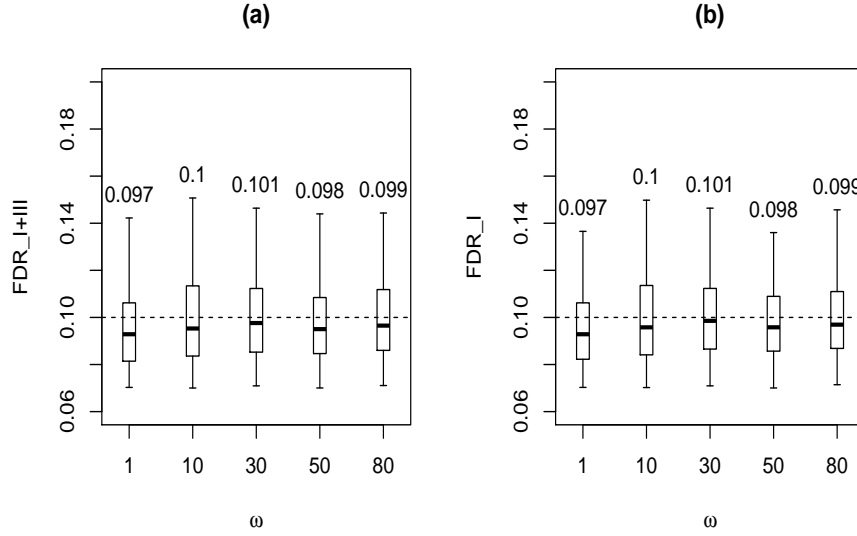


Figure 4: The boxplots of  $FDR_{I+III}$  and  $FDR_I$  based on 200 replications, respectively. The boxplots' horizontal lines are the 0.05, 0.25, 0.50, 0.75 and 0.95 quantiles of  $FDR_{I+III}$  and  $FDR_I$  versus  $\omega$ , and the numbers of above the boxplots are the means of  $FDR_{I+III}$  and  $FDR_I$ .

## 4.2 Real data analysis

To illustrate the proposed method, we analyze BLC mean correct latency for action video game players (AVGPs) and non action video game players (NAVGP). The data consists of 84 girls and 57 boys from primary and secondary schools, aging from 6 to 13 years old. They were recruited to answer the video game playing questionnaire. Using data from the questionnaire, which were collected separately from children and from their parents for verification, these 141 students were subdivided into two groups: the AVGPs group (56%) and the NAVGPs group (44%). Then, they were required to finish the Big/Little Circle test via an action video game,

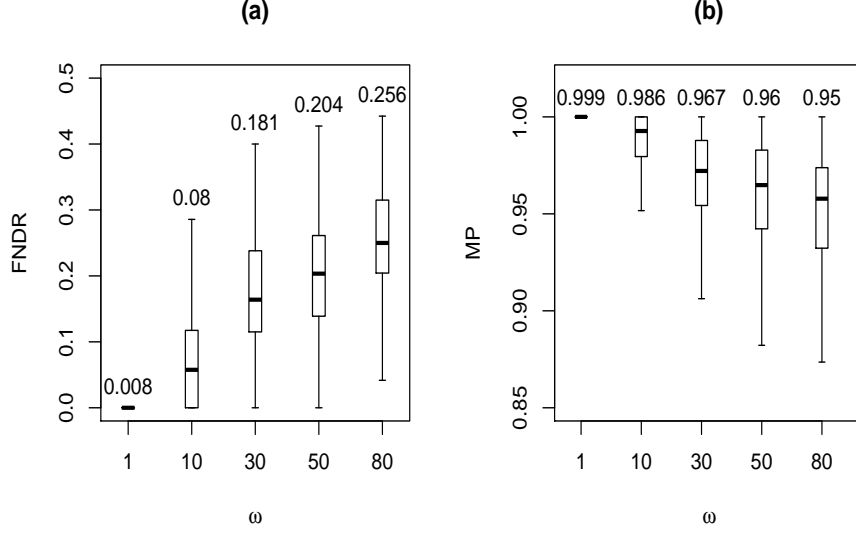


Figure 5: The boxplots of FNRD and MP based on 200 replications, respectively. The boxplots' horizontal lines are the 0.05, 0.25, 0.50, 0.75 and 0.95 quantiles of FNRD and MP versus  $\omega$ , and the numbers of above the boxplots are the means of FNRD and MP.

Table 2: True errors and averages (standard deviation) of estimated errors when  $\omega = 80$

When controlling $\text{FDR}_I$ at 0.1				
Errors	$\text{mFDR}_I$	$\text{mFDR}_{III}$	$\text{mFDR}_{I+III}$	$\text{mFNRD}$
True	0.099	0.003	0.102	0.260
Estimated	0.098	0.003	0.101	0.259
	(0.002)	(0.002)	(0.002)	(0.040)
When controlling $\text{FDR}_{I+III}$ at 0.1				
Errors	$\text{mFDR}_I$	$\text{mFDR}_{III}$	$\text{mFDR}_{I+III}$	$\text{mFNRD}$
True	0.096	0.003	0.099	0.256
Estimated	0.095	0.003	0.098	0.254
	(0.002)	(0.002)	(0.000)	(0.040)

which was defined as a video game genre that emphasizes hand-eye coordination and reaction-time. Our objective is to detect the areas of age that the significant differences between AVGPs group and NAVGPs group occur.

We use the Gaussian process regression model (1) to fit the data for each group. The estimated mean curves corresponding to AVGPs group and NAVGPs group are given in Figure 6. We can see that there are some crossings between these two curves. We first consider a test with from (2) and  $\Delta = 20$ , chosen by our collaborators in neuroscience.

We generate samples based on the conditional distribution of  $\mu_d^N$  on center points of 500

equal-length subintervals covering the age range  $[6, 13]$ , and test the hypotheses at each time point. Figure 6 shows the significant and non-significant areas detected by the proposed procedure, when controlling  $\text{FDR}_{\text{I+III}}$  at level 0.10. Aging from 6 to 9, the NAVGPs have significantly higher BLC mean correct latency than AVGPs, while after 9 years old, they have non-significant differences. It implies that the video game-based therapy may have significant effect on children with hemiplegia aging from 6 to 9 years old, while it may have limited help with of some of symptoms when they are more than 9 years old. The proposed procedure reports  $\widehat{\text{mFDR}}_{\text{I}} = 0.08$  and  $\widehat{\text{mFDR}}_{\text{III}} = 0.02$ , indicating the Type III errors account for about 20% of  $\text{mFDR}_{\text{I+III}}$ . It reports  $\widehat{\text{mFNDR}} = 0.22$ , implying that the means of BLC mean correct latency of NAVGPs and AVGPs groups could have differences larger than  $\Delta = 20$  in 22% of areas of age after 9 years old.

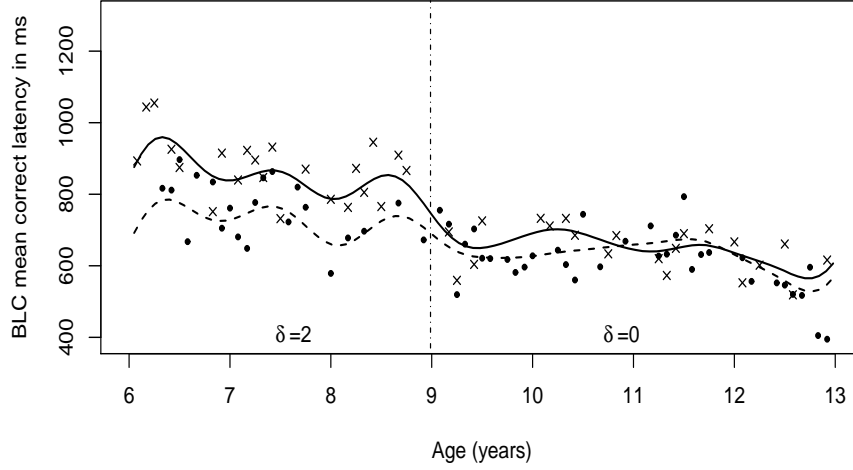


Figure 6: The significant and non-significant areas detected by the proposed procedure under  $\text{mFDR}_{\text{I+III}}$  control at 0.10. Aging from 6 to 9, the NAVGPs have significantly higher BLC mean correct latency than AVGPs ( $\delta(t) = 2$ ), while after 9 years old, they have non-significant differences ( $\delta(t) = 0$ ). The solid and the dash lines represent the estimated mean curves for the NAVGPs group ( $\times$ ) and the AVGPs group ( $\cdot$ ), respectively. The estimates of errors are:  $\widehat{\text{mFDR}}_{\text{I}} = 0.08$ ,  $\widehat{\text{mFDR}}_{\text{III}} = 0.02$ , and  $\widehat{\text{mFNDR}} = 0.22$ .

Using different values of  $\Delta$  in (2) makes the method very flexible. Figure 7 presents the results with  $\Delta = 1$  and  $\Delta = 100$ . The estimated  $\text{mFDR}_{\text{I}}$ ,  $\text{mFDR}_{\text{III}}$  and  $\text{mFNDR}$  are also calculated and presented. The former indicates there are two significant areas: one from age 6 to 9.2 and the other from 9.6 to 10.6. Consequently, with a smaller  $\Delta$   $\text{mFNDR}$  increases, i.e. there could exist 44% of areas, among declared non-significant areas, that the mean difference of these two groups is larger than  $\Delta = 1$ . The results for  $\Delta = 100$  imply that there is no detected significant area, while there would exist 30% of areas that the mean difference of two groups is



larger than  $\Delta = 100$  in whole age range [6,13]. It is not surprising that there is no rejection at all when  $\Delta \geq 100$ . This rather large number makes the result meaningless. In general, the choice of  $\Delta$  depends on a scientific question of interest.

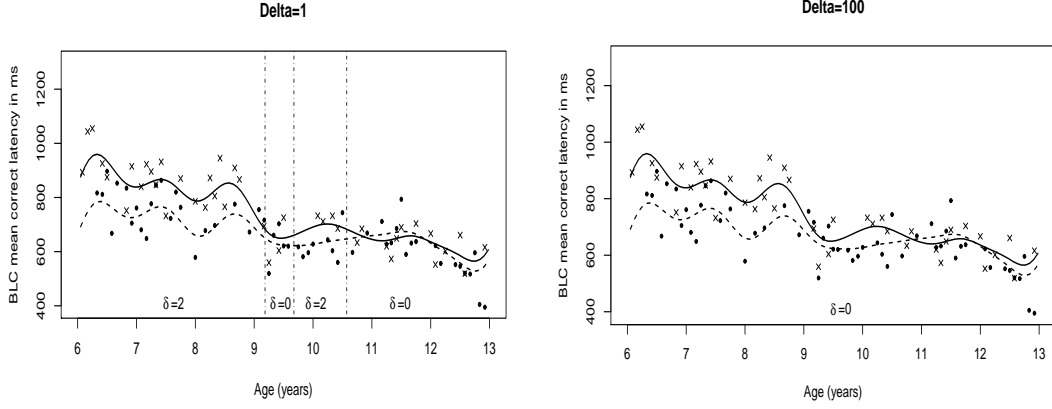


Figure 7: The significant and non-significant areas detected by the proposed procedure under  $\text{mFDR}_{\text{I+III}}$  control at 0.10. Left ( $\Delta = 1$ ): aging from 6 to 9.2 and from 9.6 to 10.6, the NAVGPs have significantly higher BLC mean correct latency than AVGPs ( $\delta(t) = 2$ ), while they have non-significant differences at other ages ( $\delta(t) = 0$ ). The estimates of errors are:  $\widehat{\text{mFDR}}_{\text{I}} = 0.01$ ,  $\widehat{\text{mFDR}}_{\text{III}} = 0.09$ , and  $\widehat{\text{mFNDR}} = 0.44$ ; right ( $\Delta = 100$ ): there is no rejection which implies that no significant area is detected. The estimates of errors are:  $\widehat{\text{mFDR}}_{\text{I}} = 0$ ,  $\widehat{\text{mFDR}}_{\text{III}} = 0$ , and  $\widehat{\text{mFNDR}} = 0.30$ .

## 5 Equivalence tests

A statistical hypothesis test is a decision rule to check whether the null hypothesis is justifiable given the observed data. We could reject the null hypothesis when there is strong evidence that it is wrong, but we could never prove it. Therefore, failure to reject  $H_0(t)$  in (2) does not mean that the difference between mean functions of two curves  $Y_1(t)$  and  $Y_2(t)$  is no more than  $\Delta$  at time  $t$ . To demonstrate similarity rather than showing differences, we sometimes need to put the similarity hypothesis into the alternative. We might thus consider the multiple testing

$$\begin{aligned} &H_{01}(t) : \mu_d(t) < -\Delta^E \quad \text{or} \quad H_{02}(t) : \mu_d(t) > \Delta^E \\ \text{versus} \quad &H_1(t) : |\mu_d(t)| \leq \Delta^E, \end{aligned} \quad (9)$$

where  $\Delta^E$  is called equivalence margin that is typically chosen as a limit below which differences are practically meaningful, and we call the test (9) the equivalence testing for functional data.

Equivalence tests have gained increasing attention during the past two decades. The goal of an equivalence test is to establish practical equivalence, which is popular used in application areas

such as medicine and biology. There are lots of procedures that have been proposed to conduct equivalence tests for scalar data. For example, Schuirmann (1987) proposed the two one-sided tests procedure for bioequivalence; Anderson and Hauck (1990) suggested the comparison of both mean and variance of the two responses when assess a generic drug's performance relative to a brand name drug; Brown et al. (1997) developed an unbiased test for the bioequivalence problem; Wang et al. (1999) discussed ways to construct a test simultaneously for all the individual pharmacokinetic parameters; Romano (2005) proposed a optimal test for testing the mean of a multivariate normal mean. Other relevant works include Chow and Liu (1992), Berger and Hsu (1996), Meyners (2012) and some of the references therein. However, in some cases the question of practical equivalence cannot be reduced to a hypothesis regarding scalar data. Recently, Fogarty and Small (2014) extended the equivalence testing framework to the functional regime. They considered an equivalence testing for overall mean difference. But they cannot test areas of the function domain with location parity. Therefore, it will be interesting to extend the proposed idea to equivalence testing (9).

Let  $z^E(t)$  be the underlying state at time  $t$ . We set  $z^E(t) = 1$  or  $2$  if hypothesis at time  $t$  is the null 1 or 2 and  $z^E(t) = 3$  if hypothesis at time  $t$  is the alternative. Let  $\delta^E(t) \in \{1, 2, 3\}$  be a decision rule for the hypothesis (9). Let  $R_k^E = \{t \in T : \delta^E(t) = k\}$  and  $V_{jk}^E = \{t \in T : z^E(t) = j, \delta^E(t) = k\}$  for  $j, k = 1, 2, 3$ . Similar to directional two-sided test (2), there also exist three types of errors in equivalence testing (9). Table 3 sums up the possible outcomes of multiple testing with two nulls. Then,  $\mathcal{L}(N_1^E) = \mathcal{L}(V_{13}^E) + \mathcal{L}(V_{23}^E)$ ,  $\mathcal{L}(N_2^E) = \mathcal{L}(V_{31}^E) + \mathcal{L}(V_{32}^E)$  and  $\mathcal{L}(N_3^E) = \mathcal{L}(V_{12}^E) + \mathcal{L}(V_{21}^E)$  are the sizes of areas corresponding to Type I, Type II and Type III errors, respectively, where  $\mathcal{L}(\cdot)$  is the Lebesgue measure on  $T$ . Hence, we define the marginal false discovery rate as  $\text{mFDR}^E = E\{\mathcal{L}(N_1^E)\}/E\{\mathcal{L}(R_3^E)\}$ , the marginal false nondiscovery rate for Type II error as  $\text{mFNDR}_{\text{II}}^E = E\{\mathcal{L}(N_2^E)\}/E\{\mathcal{L}(R_1^E \cup R_2^E)\}$  and that for Type III error as  $\text{mFNDR}_{\text{III}}^E = E\{\mathcal{L}(N_3^E)\}/E\{\mathcal{L}(R_1^E \cup R_2^E)\}$ . And for simplicity, let  $\text{mFNDR}_{\text{II+III}}^E = \text{mFNDR}_{\text{II}}^E + \text{mFNDR}_{\text{III}}^E$ .

Table 3: Outcomes of multiple testing with two nulls

	Declared as null 1 $\delta^E(t) = 1$	Declared as null 2 $\delta^E(t) = 2$	Declared as alternative $\delta^E(t) = 3$	Total
Null 1 ( $z^E(t) = 1$ )	$V_{11}^E$	$V_{12}^E$ ( <i>Type III error</i> )	$V_{13}^E$ ( <i>Type I error</i> )	$T_1^E$
Null 2 ( $z^E(t) = 2$ )	$V_{21}^E$ ( <i>Type III error</i> )	$V_{22}^E$	$V_{23}^E$ ( <i>Type I error</i> )	$T_2^E$
Alternative ( $z^E(t) = 3$ )	$V_{31}^E$ ( <i>Type II error</i> )	$V_{32}^E$ ( <i>Type II error</i> )	$V_{33}^E$	$T_3^E$
Total	$R_1^E$	$R_2^E$	$R_3^E$	$T$

We applied the equivalence testing (9) to the executive function study. The results are presented in Figure 8. The non-significant areas (i.e. the mean curves are different) obtained

by using  $\Delta^E = 140$  is similar to the ones using test (2) with  $\Delta = 20$  (see Figure 6). One reason might be the  $\text{mFDR}^E$  (analogous to Type I error) controlled here is actually the  $\text{mFNDR}$  (analogous to Type II error) in the multiple testing (2), and the  $\text{mFNDR}_{\text{II+III}}^E$  (analogous to the sum of Type II and III errors) minimized in the equivalence testing (9) is actually the  $\text{mFDR}_{\text{I+III}}$  (analogous to the sum of Type I and III errors) in test (2), which shows the clear differences between these two different types of test. As expected when  $\Delta^E \leq 100$ , there is no rejection.

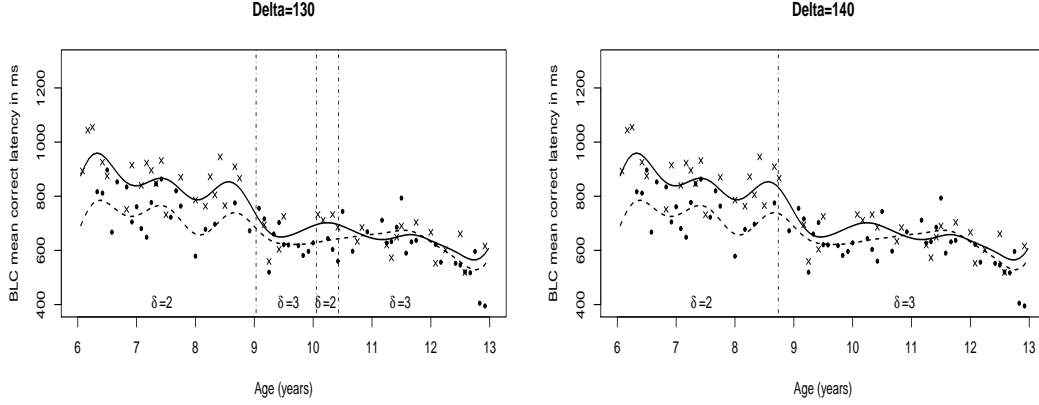


Figure 8: The equivalent and non-equivalent areas detected by the proposed procedure under  $\text{mFDR}^E$  control at 0.10. Left ( $\Delta^E = 130$ ): aging from 9 to 10 and after 10.5, the BLC mean correct latency of NAVGPs and AVGPs are similar ( $\delta^E(t) = 3$ ), while the NAVGPs have higher BLC mean correct latency than AVGPs at other ages ( $\delta^E(t) = 2$ ). The estimates of errors are:  $\widehat{\text{mFDR}}^E = 0.10$ ,  $\widehat{\text{mFNDR}}_{\text{II}}^E = 0.25$ , and  $\widehat{\text{mFNDR}}_{\text{III}}^E = 0.001$ ; Right ( $\Delta^E = 140$ ): aging after 8.8, the BLC mean correct latency of NAVGPs and AVGPs are similar ( $\delta^E(t) = 3$ ), while the NAVGPs have higher BLC mean correct latency than AVGPs from 6 to 8.8 ( $\delta^E(t) = 2$ ). The estimates of errors are:  $\widehat{\text{mFDR}}^E = 0.10$ ,  $\widehat{\text{mFNDR}}_{\text{II}}^E = 0.27$ , and  $\widehat{\text{mFNDR}}_{\text{III}}^E = 0$ .

## 6 Discussion

In this paper we proposed a method based on large scale multiple testing to detect differences of the means of two curves. It can automatically detect the significant areas and at the same time control the directional error. By taking advantage of the functional nature of the data, we introduce a nonparametric Gaussian process regression model for simultaneous two-sided tests. We are thus able to make inference at any point in a continuum and derive a procedure which optimally controls directional false discovery rates. To make it workable in practice, an approximation procedure is proposed via a finite approximation strategy. We show that the proposed procedure controls directional false discovery rates at any specified level asymptotically.

Related to the topic discussed in this paper, some interesting problems are worth further development. Though simulation studies validate the good control ability of the proposed procedure over both Type I and directional errors, the estimation of the unknown model parameters may affect the power of the testing method. It is therefore important for us to discuss the asymptotic optimality of the data-driven procedure with estimated model parameters in a more systematic fashion. And this paper focuses mainly on the problem defined in one-dimensional domain. It will be interesting to extend the idea to more complicated case, such as the problem defined in two- or three-dimensional spatial domain, or in temporal-spatio domain. Gaussian process regression model can cope with problems with multidimensional covariates. This good feature makes such extension feasible. On the other side of the spectrum, Fogarty and Small (2014) considered an equivalence testing for overall mean difference with dynamic bands. To extend the equivalence testing (9) to a more general case with dynamic lower and upper equivalence bands is another interesting direction for future investigation.

## Acknowledgements

We would like to thank Professor J. Eyre of the Institute of Neuroscience of Newcastle University in UK for allowing us to use their experimental data. Xu was supported by the Natural Science Foundation of Jiangsu Province, China (No. BK20140617). Lee was supported by the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2014M3C7A1062896).

# References

- [1] Anderson, S. and Hauck, W.W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Pharmacodynamics*, **18**, 259-273.
- [2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289-300.
- [3] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165-1188.
- [4] Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, **100**, 71-81.
- [5] Berger, R.L. and Hsu, J. (1996). Bioequivalence trials, intersection-union tests, and equivalence confidence sets (with Discussion). *Statistical Science*, **11**, 283-319.
- [6] Bjørnstad, J.F. (1996). On the generalization of the likelihood function and likelihood principle. *Journal of the American Statistical Association*, **91**, 791-806.
- [7] Bosq, D. (2000). *Linear Processes in Function Spaces*. Springer, New York.
- [8] Brown, L.D., Hwang, J.T., and Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics*, **25**, 2345-2367.
- [9] Chow, S.C. and Liu, J.P. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies*. New York: Marcel Dekker.
- [10] Clements, N., Sarkar, S.K., Zhao, Z., and Kim, D.-Y. (2014). Applying multiple testing procedures to detect change in East African vegetation. *The Annals of Applied Statistics*, **8**, 286-308.
- [11] Cox, D.D. and Lee, J.S. (2008). Pointwise testing with functional data using the Westfall-Young randomization method. *Biometrika*, **95**, 621-634.
- [12] Cuesta-Albertos, J.A. and Febrero-Bande, M. (2010). Multiway ANOVA for functional data. *TEST*, **19**, 537-557.
- [13] Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational Statistics & Data Analysis*, **47**, 111-122.
- [14] Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009), Multilevel Functional Principal Component Analysis, *The Annals of Applied Statistics*, **3**, 458-488.

- [15] Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96-104.
- [16] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, **102**, 93-103.
- [17] Estévez-Pérez, G. and Vilar, J.A. (2008). Functional anova starting from discrete data: an application to air quality data. *Environmental and Ecological Statistics*, **20**, 495-515.
- [18] Ferraty, F. and Romain, Y. (2011). *The Oxford Handbook of Functional Data Analysis*. Oxford University Press.
- [19] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York.
- [20] Fogarty, C.B. and Small, D.S. (2014). Equivalence testing for functional data with an application to comparing pulmonary function devices. *The Annals of Applied Statistics*, **8**, 2002-2026.
- [21] French, J.P. and Sain, S.R. (2013). Spatio-temporal exceedance locations and confidence regions. *The Annals of Applied Statistics*, **7**, 1421-1449.
- [22] Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, **32**, 1035-1061.
- [23] Guo, W, Sarkar, S.K., and Peddada, S.D. (2010). Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics*, **66**, 485-492.
- [24] Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer, New York.
- [25] Horváth, L., Kokoszka, P., and Reeder, R. (2013). Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society Series B*, **75**, 103-122.
- [26] Lee, Y. and Bjørnstad, J.F. (2013). Extended likelihood approach to large-scale multiple testing. *Journal of the Royal Statistical Society Series B*, **75**, 553-575.
- [27] Lee, D. and Lee, Y. (2014). Extended likelihood approach to multiple test with directional error control under hidden Markov random field model. Technical report. Department of Statistics, Seoul National University, Korea.
- [28] Lee, Y. and Nelder, J.A. (1996). Hierarchical GLMs (with discussion). *Journal of the Royal Statistical Society Series B*, **58**, 619-673.

- [29] Leventhal, L. and Huynh, C. (1996). Directional decisions for two-tailed tests: Power, error rates, and sample size. *Psychological Methods*, **1**, 278-292.
- [30] Liu, J., Zhang, C., McCarty, C., Peissig, P., Burnside, E., and Page, D. (2012). Graphical-model based multiple testing under dependence, with applications to genome-wide association studies. *The 28th Conference on Uncertainty in Artificial Intelligence*.
- [31] Meyners, M. (2012). Equivalence tests - A review. *Food Quality and Preference*, **26**, 231-245.
- [32] Moore, D. P. and Puri, B. K. (2012). *Textbook of Clinical Neuropsychiatry and Behavioral Neuroscience 3E*, CRC Press, Taylor & Francis Group.
- [33] Müller, H.-G. (2005). Functional modelling classification longitudinal. *Scandinavian Journal of Statistics*, **32**, 223-240.
- [34] Ramsay, J., Hoocher, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer, New York.
- [35] Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. 2nd edition. Springer, New York.
- [36] Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA.
- [37] Romano, J.P. (2005). Optimal testing of equivalence hypotheses. *The Annals of Statistics*, **33**, 1036-1047.
- [38] Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657-680.
- [39] Shen, Q. and Faraway, J. (2004). An F test for linear models with functional responses. *Statistica Sinica*, **14**, 1239-1257.
- [40] Shi, J. and Choi, T. (2011). *Gaussian Process Regression Analysis for Functional Data*. Chapman and Hall/CRC, London.
- [41] Staicu, A.-M., Li, Y., Crainiceanu, C.M., and Ruppert, D. (2014). Likelihood ratio tests for dependent data with application to longitudinal and functional data analysis. *Scandinavian Journal of Statistics*, **41**, 932-949.
- [42] Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, **64**, 479-498.
- [43] Sun, W. and Cai, T.T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society Series B*, **71**, 393-424.

- [44] Sun, W., Reich, B., Cai, T.T., Guindani, M., and Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society Series B*, **77**, 59-83.
- [45] Wang, B. and Shi, J. (2014). Generalized Gaussian process regression model for non-Gaussian functional data. *Journal of the American Statistical Association*, **109**, 1123-1133.
- [46] Wang, W., Hwang, J.T.G., and Dasgupta, A. (1999). Statistical tests for multivariate bioequivalence. *Biometrika*, **86**, 395-402.
- [47] Yao, F., Müller, H.-G., and Wang, J. L. (2005), Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, **100**, 577-590.
- [48] Zhang, C., Fan, J., and Yu, T. (2011). Multiple testing via  $FDR_L$  for large-scale imaging data. *The Annals of Statistics*, **39**, 613-642.
- [49] Zhang, J.-T., Liang, X., and Xiao, S. (2010). On the two-sample behrens-fisher problem for functional data. *Journal of Statistical Theory and Practice*, **4**, 571-587.



# Appendix

## Appendix A. Technical Proofs

**Proof of Theorem 1.** We first prove Theorem 1(1). If  $\lambda_2 = 1$  and  $\lambda_1 = \lambda_3 = \lambda$ , the loss function (3) becomes

$$L(\delta, z; \lambda) = \mathcal{L}(N_2) + \lambda\{\mathcal{L}(N_1) + \mathcal{L}(N_3)\},$$

which can be re-written as

$$\begin{aligned} L(\delta, z; \lambda) &= \sum_{k=1}^2 \int_T I(z(t) = k) I(\delta(t) = 0) d\mathcal{L}(t) + \lambda \left\{ \sum_{k=1}^2 \int_T I(z(t) = 0) I(\delta(t) = k) d\mathcal{L}(t) \right. \\ &\quad \left. + \sum_{j=1}^2 \sum_{k \neq 0, j} \int_T I(z(t) = k) I(\delta(t) = j) d\mathcal{L}(t) \right\}. \end{aligned}$$

Then, the posterior classification risk is

$$\begin{aligned} \mathbb{E}\{L(\delta, z; \lambda) \mid \mathcal{D}\} &= \sum_{k=1}^2 \int_T I(\delta(t) = 0) \mathbb{P}(z(t) = k \mid \mathcal{D}) d\mathcal{L}(t) + \lambda \left\{ \sum_{k=1}^2 \int_T I(\delta(t) = k) \right. \\ &\quad \left. \mathbb{P}(z(t) = 0 \mid \mathcal{D}) d\mathcal{L}(t) + \sum_{j=1}^2 \sum_{k \neq 0, j} \int_T I(\delta(t) = j) \mathbb{P}(z(t) = k \mid \mathcal{D}) d\mathcal{L}(t) \right\} \\ &= \int_T \left\{ I(\delta(t) = 0) \mathbb{P}(z(t) \neq 0 \mid \mathcal{D}) + \lambda \sum_{k=1}^2 I(\delta(t) = k) \mathbb{P}(z(t) \neq k \mid \mathcal{D}) \right\} d\mathcal{L}(t) \\ &= \int_T \mathbb{P}(z(s) \neq 0 \mid \mathcal{D}) \left\{ I(\delta(t) = 0) + \sum_{k=1}^2 I(\delta(t) = k) \frac{\lambda \mathbb{P}(z(t) \neq k \mid \mathcal{D})}{\mathbb{P}(z(s) \neq 0 \mid \mathcal{D})} \right\} d\mathcal{L}(t). \end{aligned}$$

Therefore, the optimal decision rule  $\delta^{(I+III)} = \{\delta^{(I+III)}(t) : t \in T\} = \operatorname{argmin}_{\delta} \mathbb{E}\{L(\delta, z; \lambda) \mid \mathcal{D}\}$  is

$$\begin{aligned} \delta^{(I+III)}(t) &= k \text{ if } \frac{\mathbb{P}(z(t) \neq 0 \mid \mathcal{D})}{\mathbb{P}(z(t) \neq k \mid \mathcal{D})} > \lambda \text{ and } \mathbb{P}(z(t) = k \mid \mathcal{D}) = \max_{j=1,2} \mathbb{P}(z(t) = j \mid \mathcal{D}), \\ &= 0 \text{ otherwise,} \end{aligned}$$

which finishes the proof of Theorem 1(1). Similar arguments can be used to prove Theorem 1(2).

□

**Proof of Theorem 2.** Given an  $\text{mFDR}_{I+III}$  level  $\alpha$ , consider a decision rule  $\delta = \{\delta(t) : t \in T\}$  with  $\text{mFDR}_{I+III}\{\delta\} \leq \alpha$ . Let  $R$  be the expected rejection area for  $\delta$ . Define  $\Upsilon(t) = \{\min_{1 \leq k \leq 2} \mathbb{P}(z(t) = k \mid \mathcal{D}) + \mathbb{P}(z(t) = 0 \mid \mathcal{D})\} / \mathbb{P}(z(t) \neq 0 \mid \mathcal{D})$ . Then, according to the definition of  $\delta^{(I+III)}$ , its corresponding expected rejection area is

$$R(\lambda) = \mathbb{E} \int_T I(\Upsilon(t) \leq \lambda^{-1}) d\mathcal{L}(t) = \int_T \mathbb{P}(\Upsilon(t) \leq \lambda^{-1}) d\mathcal{L}(t).$$

Hence,  $R(\lambda)$  is decreasing with  $\lambda$ . In addition, it is easy to see that

$$\lim_{\lambda \rightarrow 0} \frac{R(\lambda)}{\mathcal{L}(T)} = 1, \text{ and } \lim_{\lambda \rightarrow \infty} R(\lambda) = 0.$$

Consequently, for a given expected rejection area  $R$  determined by  $\delta$ , there exists a unique  $\lambda(R)$  such that the decision rule  $\delta^{(I+III)}$  has the same expected rejection area.

Further, for  $\delta^{(I+III)}$ , define  $\text{TD}_{\delta^{(I+III)}}$ ,  $\text{FD}_{\delta^{(I+III)I}}$  and  $\text{FD}_{\delta^{(I+III)III}}$  as the expected true discovery area, expected false discovery area related to Type I error and expected false discovery area related to Type III error, respectively. Then, we have

$$\begin{aligned}\text{TD}_{\delta^{(I+III)}} &= \sum_{k=1}^2 \mathbb{E} \int_T I(z(t) = k) I(\delta^{(I+III)}(t) = k) d\mathcal{L}(t), \\ \text{FD}_{\delta^{(I+III)I}} &= \sum_{k=1}^2 \mathbb{E} \int_T I(z(t) = 0) I(\delta^{(I+III)}(t) = k) d\mathcal{L}(t), \\ \text{FD}_{\delta^{(I+III)III}} &= \sum_{j=1}^2 \sum_{k \neq 0, j} \mathbb{E} \int_T I(z(t) = k) I(\delta^{(I+III)}(t) = j) d\mathcal{L}(t),\end{aligned}$$

and  $R(\lambda) = \text{TD}_{\delta^{(I+III)}} + \text{FD}_{\delta^{(I+III)I}} + \text{FD}_{\delta^{(I+III)III}}$ . Similarly, let  $\text{TD}_{\delta}$ ,  $\text{FD}_{\delta I}$  and  $\text{FD}_{\delta III}$  be the expected true discovery area, expected false discovery area related to Type I error and expected false discovery area related to Type III error for  $\delta$ , respectively. Then, it also holds that  $R(\lambda) = \text{TD}_{\delta} + \text{FD}_{\delta I} + \text{FD}_{\delta III}$ . For  $\zeta = \delta^{(I+III)}$ ,  $\delta$ , consider the loss function

$$\begin{aligned}L(z, \zeta) &= \mathcal{L}(N_2) + \lambda \{ \mathcal{L}(N_1) + \mathcal{L}(N_3) \} \\ &= \sum_{k=1}^2 \int_T I(z(t) = k) I(\zeta(t) = 0) d\mathcal{L}(t) + \lambda \left\{ \sum_{k=1}^2 \int_T I(z(t) = 0) I(\zeta(t) = k) d\mathcal{L}(t) \right. \\ &\quad \left. + \sum_{j=1}^2 \sum_{k \neq 0, j} \int_T I(z(t) = k) I(\zeta(t) = j) d\mathcal{L}(t) \right\}.\end{aligned}$$

Then, the risk for  $\delta$  and  $\delta^{(I+III)}$  is

$$\begin{aligned}\text{EL}(z, \zeta) &= \sum_{k=1}^2 \mathbb{E} \int_T I(z(t) = k) \{1 - I(\zeta(t) \neq 0)\} d\mathcal{L}(t) + \lambda (\text{FD}_{\zeta I} + \text{FD}_{\zeta III}) \\ &= \int_T \sum_{k=1}^2 \mathbb{P}(z(t) = k) d\mathcal{L}(t) - \mathbb{E} \int_T \sum_{k=1}^2 I(z(t) = k) I(\zeta(t) = k) d\mathcal{L}(t) \\ &\quad - \mathbb{E} \int_T \sum_{j=1}^2 \sum_{k \neq 0, j} I(z(t) = k) I(\zeta(t) = j) d\mathcal{L}(t) + \lambda (\text{FD}_{\zeta I} + \text{FD}_{\zeta III}) \\ &= \int_T \sum_{k=1}^2 \mathbb{P}(z(t) = k) d\mathcal{L}(t) + \lambda (\text{FD}_{\zeta I} + \text{FD}_{\zeta III}) - (\text{TD}_{\zeta} + \text{FD}_{\zeta III}).\end{aligned}$$

Since  $\text{EL}(z, \delta^{(I+III)}) \leq \text{EL}(z, \delta)$ , it implies that  $\text{FD}_{\delta^{(I+III)I}} + \text{FD}_{\delta^{(I+III)III}} \leq \text{FD}_{\delta I} + \text{FD}_{\delta III}$  and  $\text{TD}_{\delta^{(I+III)}} + \text{FD}_{\delta^{(I+III)III}} \geq \text{TD}_{\delta} + \text{FD}_{\delta III}$ . Therefore,

$$\text{mFDR}_{I+III}\{\delta^{(I+III)}\} = \frac{\text{FD}_{\delta^{(I+III)I}} + \text{FD}_{\delta^{(I+III)III}}}{R(\lambda)} \leq \frac{\text{FD}_{\delta I} + \text{FD}_{\delta III}}{R(\lambda)} = \text{mFDR}_{I+III}\{\delta\} \leq \alpha,$$

and

$$\text{mFNDNR}\{\delta^{(I+III)}\} = \frac{\text{TD}_{\delta^{(I+III)}} + \text{FD}_{\delta^{(I+III)}}}{\mathcal{L}(T) - R(\lambda)} \leq \frac{\text{TD}_{\delta} + \text{FD}_{\delta}}{\mathcal{L}(T) - R(\lambda)} \leq \text{mFNDNR}\{\delta\}.$$

□

To prove the procedure (5) is asymptotically valid for  $\text{FDR}_I$  control, we first need the following regularity conditions.

C1 Let  $\rho > 0$  be a small positive constant. For  $\mu_0 = -\Delta$  or  $\Delta$ ,  $\int_T \mathbb{P}(|\mu_d(t) - \mu_0| < \rho) d\mathcal{L}(t) \rightarrow 0$  as  $\rho \rightarrow 0$ .

C2 Let  $\mu_d^N(t) = \sum_{i=1}^N \mu_d(t_i^*) I(s_{i-1} \leq t < s_i)$ . Assume the sequence of partitions  $\{\cup_{i=1}^N [s_{i-1}, s_i) : N = 1, 2, \dots\}$  satisfies that for any given  $\rho > 0$ ,  $\int_T \mathbb{P}(|\mu_d(t) - \mu_d^N(t)| \geq \rho) d\mathcal{L}(t) \rightarrow 0$  as  $N \rightarrow \infty$ .

Conditions C1 and C2 are similar to conditions 1-2 in Sun et al. (2015). Condition C1 states that  $\{\mu_d(t) : t \in T\}$  is a smooth process that does not degenerate at both points  $-\Delta$  and  $\Delta$ . It is naturally holds when  $\{\mu_d(t) : t \in T\}$  is a continuous random process, which ensures that the inequality between  $z(t)$  and  $z^N(t)$  only occurs with a small chance when  $|\mu_d^N(t) - \mu_d(t)|$  is small, where  $z^N(t) = \sum_{i=1}^N z(t_i^*) I(s_{i-1} \leq t < s_i)$ . Condition C2 requires that the partition  $T = \cup_{i=1}^N [s_{i-1}, s_i)$  should produce roughly homogeneous subintervals so that the decision at the center point  $t_i^*$  can be a good representation of the decision process on subinterval  $[s_{i-1}, s_i)$ . Then, we will need a lemma of Sun et al. (2015) (Lemma 2). We re-state the result.

LEMMA 6.1. Under conditions C1 and C2,  $\lim_{N \rightarrow \infty} \int_T \mathbb{P}(z(t) \neq z^N(t)) d\mathcal{L}(t) = 0$ .

**Proof of Theorem 3.** Let  $S_k(t) = \mathbb{P}(z(t) = k \mid \mathcal{D})$ ,  $k = 0, 1, 2$ . According to the definition of  $\text{FDR}_I$ , the  $\text{FDR}_I$  level of procedure (5) is

$$\begin{aligned} \text{FDR}_I &\leq \mathbb{E} \left\{ \frac{1}{\mathcal{L}(R_1 \cup R_2) \vee 1} \int_0^1 S_0(t) I(\delta^{(I)}(t) \neq 0) d\mathcal{L}(t) \right\} \\ &= \mathbb{E} \left\{ \frac{1}{\mathcal{L}(R_1 \cup R_2) \vee 1} \sum_{i=1}^N I(\delta^{(I)}(t_i^*) \neq 0) \int_{s_{i-1}}^{s_i} S_0(t) d\mathcal{L}(t) \right\} \\ &= \mathbb{E} \left\{ \frac{1}{N(\mathcal{L}(R_1 \cup R_2) \vee 1)} \sum_{i=1}^N I(\delta^{(I)}(t_i^*) \neq 0) S_0(t_i^*) \right\} + A_N, \end{aligned}$$

where  $A_N = \mathbb{E}[\{\mathcal{L}(R_1 \cup R_2) \vee 1\}^{-1} \sum_{i=1}^N I(\delta^{(I)}(t_i^*) \neq 0) \int_{s_{i-1}}^{s_i} (S_0(t_i^*) - S_0(t)) d\mathcal{L}(t)]$ .

Further, let  $S_k^N(t) = \mathbb{P}(z^N(t) = k \mid \mathcal{D})$ ,  $k = 0, 1, 2$ . Note that  $\mathbb{E}|S_k^N(t) - S_k(t)| = \mathbb{P}(z^N(t) \neq k, z(t) \neq k) + \mathbb{P}(z(t) = k, z^N(t) \neq k)$ . Then, an application of Lemma 6.1 yields that

$$\begin{aligned} A_N &= \mathbb{E} \left\{ \frac{1}{\mathcal{L}(R_1 \cup R_2) \vee 1} \int_0^1 I(\delta^{(I)}(t) \neq 0) (S_0^N(t) - S_0(t)) d\mathcal{L}(t) \right\} \\ &\leq \int_0^1 \mathbb{E}[I(\delta^{(I)}(t) \neq 0) \{S_0^N(t) - S_0(t)\}] d\mathcal{L}(t) \\ &\leq 2 \int_0^1 \mathbb{P}(z(t) \neq z^N(t)) d\mathcal{L}(t) \rightarrow 0, \end{aligned}$$

where the second inequality follows from the fact that  $\{\mathcal{L}(R_1 \cup R_2) \vee 1\}^{-1} \leq 1$ . Since the proposed procedure guarantees that

$$\frac{1}{N(\mathcal{L}(R_1 \cup R_2) \vee 1)} \sum_{i=1}^N I(\delta^{(I)}(t_i^*) \neq 0) S_0(t_i^*) \leq \alpha$$

for all realization of  $\mathcal{D}$ , the  $\text{FDR}_I$  is controlled at level  $\alpha$  asymptotically.  $\square$

## Appendix B. Derivation of equations (8)

Note that  $f_{\Theta}(Y, \tilde{\mu}, \tilde{\mu}_d) = f_{\Theta}(Y) f_{\Theta}(\tilde{\mu}, \tilde{\mu}_d \mid \mathcal{D})$ , where  $f_{\Theta}(Y)$  does not contain any information about  $\tilde{\mu}$  and  $\tilde{\mu}_d$ . Hence, we have

$$\begin{aligned} f_{\Theta}(\tilde{\mu}, \tilde{\mu}_d \mid \mathcal{D}) &\propto f_{\Theta}(Y, \tilde{\mu}, \tilde{\mu}_d) \\ &\propto \phi(\tilde{\mu} \mid \mathbf{0}, \mathbf{K}_n) \phi(\tilde{\mu}_d \mid \mathbf{0}, \mathbf{\Gamma}_{n1}) \prod_{i=1}^{n_1} \phi(Y_{1i} \mid \mu(t_{1i}) + \mu_d(t_{1i}), \sigma^2) \prod_{i=1}^{n_2} \phi(Y_{2i} \mid \mu(t_{2i}), \sigma^2). \end{aligned}$$

Then, it is straightforward to know that

$$\begin{aligned} f_{\Theta}(\tilde{\mu}_d \mid \mathcal{D}) &= \int f_{\Theta}(\tilde{\mu}, \tilde{\mu}_d \mid \mathcal{D}) d\tilde{\mu} \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} (\tilde{\mu}_d - \mathbf{A}^{-1}\mathbf{b})^T \mathbf{A} (\tilde{\mu}_d - \mathbf{A}^{-1}\mathbf{b}) \right\}, \end{aligned}$$

where  $\mathbf{A} = \sigma^2 \mathbf{\Gamma}_{n1}^{-1} + \mathbf{I}_{n1} - \mathbf{\Sigma}_{11}$  and  $\mathbf{b} = (\mathbf{I}_{n1} - \mathbf{\Sigma}_{11})\mathbf{Y}_1 - \mathbf{\Sigma}_{12}\mathbf{Y}_2$ . It implies that  $\tilde{\mu}_d = \mathbf{A}^{-1}\mathbf{b} + \epsilon_1$  with  $\epsilon_1 \sim N(\mathbf{0}, \sigma^2 \mathbf{A}^{-1})$ . On the other hand, note that  $(\tilde{\mu}_d^T, \mu_d^{NT})^T$  follows a multivariate normal distribution with mean zero and covariance matrix  $\mathbf{\Gamma}$ , where

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_{n1} & \mathbf{\Psi}^T(T^*) \\ \mathbf{\Psi}(T^*) & \mathbf{\Gamma}_N \end{pmatrix}.$$

Thus, we have  $\mu_d^N = \mathbf{\Psi}(T^*)\mathbf{\Gamma}_{n1}^{-1}\tilde{\mu}_d + \epsilon_2$  with  $\epsilon_2 \sim N(\mathbf{0}, \mathbf{\Gamma}_N - \mathbf{\Psi}(T^*)\mathbf{\Gamma}_{n1}^{-1}\mathbf{\Psi}^T(T^*))$ . Consequently,  $\mu_d^N = \mathbf{\Psi}(T^*)\mathbf{\Gamma}_{n1}^{-1}\mathbf{A}^{-1}\mathbf{b} + \mathbf{\Psi}(T^*)\mathbf{\Gamma}_{n1}^{-1}\epsilon_1 + \epsilon_2$ , so the conditional distribution of  $\mu_d^N$  given  $\mathcal{D}$  is a multivariate normal distribution with mean  $\mathbf{\Psi}(T^*)\mathbf{\Gamma}_{n1}^{-1}\mathbf{A}^{-1}\mathbf{b}$  and covariance matrix  $\sigma^2 \mathbf{\Psi}(T^*)\mathbf{\Gamma}_{n1}^{-1}\mathbf{A}^{-1}\mathbf{\Gamma}_{n1}^{-1}\mathbf{\Psi}^T(T^*) + \mathbf{\Gamma}_N - \mathbf{\Psi}(T^*)\mathbf{\Gamma}_{n1}^{-1}\mathbf{\Psi}^T(T^*)$ , i.e.,  $\mu_d^N \mid \mathcal{D} \sim N(\tilde{\mu}, \mathbf{\Lambda})$ .